DOCUMENT RESUME

ED 420 722 TM 028 444

AUTHOR Chang, Shun-Wen; Twu, Bor-Yaun
TITLE A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing.
PUB DATE 1998-04-00
NOTE 52p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Adaptive Testing; Algorithms; Comparative Analysis; *Computer Assisted Testing; Estimation (Mathematics); Item Banks; Monte Carlo Methods; Simulation; *Test Items
IDENTIFIERS *Item Exposure (Tests); *Test Security

ABSTRACT

This study investigated and compared the properties of five methods of item exposure control within the purview of estimating examinees' abilities in a computerized adaptive testing (CAT) context. Each of the exposure control algorithms was incorporated into the item selection procedure and the adaptive testing progressed based on the CAT design established for this study. Using Monte Carlo simulation, the merits and shortcomings of these strategies were considered under different item pool sizes and different maximum exposure rates and were evaluated in light of test security, item overlap rate, and the conditional standard errors of measurement. The ultimate goal of the study was to provide guidelines for choosing appropriate methods of controlling item exposure so that test security concern in the CAT environment could be lessened. Among the five exposure control algorithms investigated in this study, the conditional procedure developed by M. Stocking and C. Lewis (1995) best served the purpose of controlling the observed exposure rates to the desired values as well as producing the lowest test overlap rates. The trade-off was that measurement precision was sacrificed to some extent. (Contains 4 tables, 9 figures, and 28 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 420 722

A Comparative Study of Item Exposure Control Methods in Computerized Adaptive Testing

Shun-Wen Chang and Bor-Yaun Twu

The University of Iowa

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Shun-Wen Chang

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- ✓ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- • Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028444

Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, April 1998

Abstract

This study investigated and compared the properties of five methods of item exposure control within the purview of estimating examinees' abilities in a CAT context. Each of the exposure control algorithms was incorporated into the item selection procedure and the adaptive testing progressed based on the CAT design established for this study. The merits and shortcomings of these strategies were considered under different item pool sizes and different desired maximum exposure rates and were evaluated in light of test security, item overlap rate, and the conditional standard errors of measurement.

The ultimate goal of this study was to provide guidelines for choosing appropriate methods of controlling item exposure so that the test security concern in the CAT environment could be lessened.

Acknowledgement

The authors wish to thank Timothy Ansley and Deborah Harris for their assistance and comments.

The development of computerized adaptive tests (CATs) and the research on computerized adaptive testing (CAT) have reached unprecedented levels. With a great deal of effort over the last two decades, theoretical hurdles in implementing large-scale CATs have been gradually overcome. The remarkable progress has made CAT an operational testing practice at present.

With a growing number of large-scale applications in CAT, measurement professionals are encountering many practical issues while the theory is translated into practice. One issue arises when some items appear repeatedly within a short period of time or within the same geographical area. According to current item selection procedures, items yielding the highest information values for a given provisional ability estimate are selected often. This results in high measurement precision and administrative efficiency. If an item is presented too often, however, a large proportion of test-takers will see this item. The problem arrives when some test-takers have access to the questions before test administrations. Frequently appearing test items may soon be compromised, causing substantial concern for the makers of high stakes tests. A high rate of *item exposure* leads to a large test security risk.

Controlling item exposure is one of the practical issues that must now be addressed (Davey & Parshall, 1995; Featherman, Subhiyah, & Hadadi, 1996; Hetter & Sympson, 1995; Mills & Stocking, 1996; Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995; Stocking, 1993; Stocking & Lewis, 1995a, 1995b; Way, 1997). Due to the opportunity that candidates might have to communicate with people who have recently taken CATs using the same item pool, Stocking (1993) pointed out that for CAT to be a serious competitor to traditional paper-and-pencil (P&P) testing, methods must be developed to limit the exposure of items in order to ensure fairness to all examinees, as is currently the case for traditional tests. Schaeffer et al. (1995) claimed that the question of the most effective way to diminish item exposure must be answered, not to mention the fact that even a low exposure rate could lead to a high exposure volume after some time.

To date, there have been several algorithms proposed to control item exposure rates (Davey & Parshall, 1995; McBride & Martin, 1983; Stocking & Lewis, 1995a, 1995b; Sympson & Hetter, 1985). Although a few studies have compared some of the item exposure control strategies (Davey & Parshall, 1995; Featherman et al., 1996; Stocking & Lewis, 1995b), no single study has systematically examined the behavior of these approaches. No single study has provided a comprehensive understanding of the overall merits and shortcomings of these different tactics. Given the vital role of item selection in CAT implementation and the urgent calls for test security, a thorough comparative study is needed at this time.

Some Background

In the early stages of item exposure rate research, McBride and Martin (1983) attempted to increase item security by indirectly reducing the frequency with which an item appears. The 5-4-3-

2-1 algorithm they developed uses a randomization scheme to prevent the overexposure of initial items. For an examinee with an initial ability estimate, the first item to be actually administered is randomly chosen from a group of the best five items identified to be approximately equal in optimality. The second item is randomly selected from a group of the best four at the updated ability level for the examinee. Then, the third is randomly chosen from a group of the best three, the fourth is randomly chosen from a group of the best two, and the fifth and subsequent items are as selected to be optimal at that examinee's current updated ability level. The most informative item at a current ability estimate in the early testing process is thus not always administered.

Several algorithms have been proposed to directly control item exposure rates (Davey & Parshall, 1995; Stocking & Lewis, 1995a, 1995b; Sympson & Hetter, 1985). By embedding statistical mechanisms in the item selection procedure, these methods seek to control the exposure rates of items to a desired maximum value, r, that is specified in advance of testing.

The Sympson and Hetter (1985) procedure employs an exposure control parameter for each item in the pool, which is determined after a series of iterative simulations. Given that an item has been selected, whether to administer this item to the examinee depends upon the exposure control parameter of this item. For very popular items, the exposure control parameters could be as low as the pre-specified desired exposure rate, indicating that these items cannot be freely administered when they are selected. For items rarely appearing, the associated exposure control parameters could be as high as 1.0, meaning that once these items are selected, they are almost always presented. The probabilistic model of Sympson and Hetter seeks to achieve the goal that there is no item administered to more than a pre-specified fraction of examinees.

Below are the steps in this strategy (Hetter & Sympson, 1995; Sympson & Hetter, 1985):

1. Set the initial exposure control parameter K_{i_0}, one for each item, to be 1.0.
2. For a current examinee's ability estimate, select the item that is the most informative.
3. Generate a pseudo-random number, x, from a uniform distribution between 0 and 1.
4. Compare this random number x with the exposure control parameter for the selected item. Administer the item only when x is less than or equal to the item's K_{i_0}; otherwise, set the item aside and identify the next most optimal item and repeat the procedure until an item is found that can be administered. The set-aside items are excluded from the pool of remaining items for this examinee.
5. Continue the administration until the whole sample, $N(E)$, have been tested. Record the number of times each item is selected, $N(S)$, and the number of times it is administered, $N(A)$. Compute the probability that an item is selected as optimal, $P(S)$, and the probability that an item is administered, $P(A)$, by the formulas:

$$P(S) = N(S) / N(E), \text{ and } P(A) = N(A) / N(E).$$

6. Obtain a new K_{i_1} for each item:

   If $P(S) > r$, then new $K_{i_1} = r / P(S)$;
   If $P(S) \leq r$, then new $K_{i_1} = 1.0$.

7 Repeat steps 2 through 6 given the new K_{i_1} until the largest observed value of the $P(A)$ across the entire item pool is approximately equal to the pre-specified exposure rate r and the K_i for all items have stabilized in subsequent iterative simulations. The value of K_i, one for each item, obtained at the final round of iterations is the exposure control parameter to be used in real testing for item i.

The Davey and Parshall (1995) methodology provides an exposure parameter for each item that is conditioned on all other items previously administered to the examinee. In addition to restricting the exposure probabilities of items in an adaptive test, this method is proposed to minimize the extent to which sets or pairs of items appear together. It was concluded in Davey and Parshall (1995) that this procedure successfully reduces the extent to which the items overlap across tests administered for examinees with similar ability and for examinees of differing ability.

To utilize the Davey and Parshall method, an exposure table needs to be prepared. Diagonal elements of the table indicate the probability limits with which individual items can be administered given selection. These values will be small if the corresponding items tend to be selected frequently. The off-diagonal elements represent the probability limits with which a pair or set of items can appear together given selection. Similarly, the off-diagonal values will be small if the pairs of items tend to occur together very often.

A series of simulations is carried out to determine the entries of the exposure table. New entries for the diagonals of the exposure table are computed based on the pre-specified desired upper limit on the frequency. If an item appears more often than the upper limit allowed, the current diagonal value is decreased by being multiplied by 0.95. If an item occurs less often than it could be, the current diagonal value is increased by multiplying 1.04, with an upper bound value of 1.0. To calculate the new entries for the off-diagonal elements, the decision rule relies on some chi-square statistics obtained from contingency tables for pairs of items. Similarly, the off-diagonal value is adjusted by being multiplied by 0.95 or 1.04, depending on the chi-square value observed. The entire process is described in greater detail by Davey and Parshall (1995).

The Stocking and Lewis unconditional multinomial procedure (1995a) was derived by remodeling the Sympson and Hetter approach. This method develops an exposure control parameter for each item following the Sympson and Hetter algorithm but it differs in the way an

item is selected. Rather than use optimal item selection, Stocking and Lewis employed a multinomial model to select the next item for administration.

At each ability level, a list of items from most desirable to least desirable is ordered based on an algorithm such as the Stocking/Swanson weighted deviations model (WDM) (Stocking & Swanson, 1993; Swanson & Stocking, 1993) or any other method of ordering the desirability. To select an item for administration, the multinomial model is utilized as follows (Stocking & Lewis, 1995a, 1995b):

1. Establish the operant probability of administration given selection, $K^*_{i_a}$, for each item in the ordered list by the formula:

$$K^*_{i_a} = \left\{ \prod_{r=1}^{i-1} \left(1 - K_{i_a}\right) \right\} \times K_{i_a},$$

where K_{i_a} is the Sympson and Hetter exposure control parameter for item i, and $K^*_{i_a}$ is the operant probability of item i, representing the joint probability that all items before item i are rejected given selection and that item i is administered given it is selected.

The sum of these operant probabilities over all items must be equal to one so that the next item can always be found for administration. In some cases where they do not sum to one, each operant probability is adjusted by being divided by the sum.

2. Form a cumulative multinomial distribution by successive addition of the adjusted operant probabilities.
3. Generate a random number uniformly distributed between 0 and 1 and place it in the cumulative multinomial distribution to determine an item to be administered. All items appearing in the ordered list preceding the item actually administered are removed from further consideration in this adaptive test.

Stocking and Lewis (1995b) also proposed the conditional multinomial method to directly control the item exposure to examinees at the same or similar levels of proficiency. Unlike the unconditional methods where an exposure control parameter is developed for an item in order to limit the item's overall appearance in reference to the examinee group, the Stocking and Lewis conditional multinomial procedure derives for each item an exposure control parameter with respect to a particular level of examinee ability. Accordingly, this procedure results in different exposure control parameters for each item to be applied to various ability levels.

Each of the five methods introduced above seems to have some potential for controlling item exposure to some degree. While the procedures were being developed, different CAT designs were employed that varied in terms of item pool size, item selection rule, expected maximum

exposure rate and termination rule, and so forth. In order to understand the behavior of the exposure control methods in relation to each other, an investigation under a common CAT design is needed. In addition, it is logically apparent that for item pools of the same quality, a large pool allows more choices for item selection than a small pool. Also, given a sufficient number of items in the pool, a pre-specified low desired exposure rate is expected to limit the observed frequencies of item appearance to a greater extent than a pre-specified high rate. Therefore, the size of the item pool and the desired maximum rate of item exposure play decisive roles in how often items are exposed. The performance of the exposure control strategies with respect to these two factors should be understood.

The Purpose of the Present Research

The purpose of this study was to investigate and compare the properties of various methods of item exposure control within the purview of estimating examinees' abilities in a CAT context. Monte Carlo simulations were employed to carry out this research.

Under the conditions of different item pool sizes and different desired maximum exposure rates, the effectiveness and psychometric properties of the various exposure control methods were evaluated in terms of test security, item overlap rate, and the conditional standard error of measurement (CSEM). Specifically, the present study attempted to achieve the following objectives:

1. to offer more information about the properties of the various exposure control strategies;
2. to provide information about how the exposure control methods are affected by using different sizes of item pools and by imposing different upper probability limits of item exposure;
3. to provide guidelines for choosing an appropriate exposure control method in practically meeting the purposes and needs of testing programs.

DESIGN OF THE STUDY

Specification of Factors Included in the Study

I. Item Exposure Control Methods

Method of controlling item exposure rates was the primary independent variable examined in this study. Listed below are the five best available exposure control strategies which were considered for all combinations of the other factors:

(1). the 5-4-3-2-1 randomization algorithm (McBride & Martin, 1983),

(2). the Sympson and Hetter procedure (Sympson & Hetter, 1985),

(3). the Davey and Parshall methodology (Davey & Parshall, 1995),

(4). the Stocking and Lewis unconditional multinomial procedure (Stocking & Lewis, 1995a), and

(5). the Stocking and Lewis conditional multinomial procedure (Stocking & Lewis, 1995b).

The item selection with no control exercised over item exposure rates was included as part of the study for comparison purposes.

II. Item Pool Sizes

Two item pools of 360 and 720 discrete items were used in the study, with the larger item pool subsuming the smaller one. The sizes of 360 and 720 were chosen because they were from multiple forms of the ACT Assessment Mathematics Test (ACT-Math) (ACT, 1997) where one form contains 60 multiple-choice items and, most importantly, because they were acceptable pool sizes according to the CAT literature. Based on the rule of thumb suggested by Stocking (1994) that in order to support a fixed-length adaptive test of roughly one-half the length of the parallel linear form, it is necessary that the CAT item pool contains at least six to eight typical linear forms. As will be addressed later in the discussion of the test termination rule, the adaptive test in the current study was terminated after an examinee was administered 30 items. Applying Stocking's rule, the small item pool consisting of 360 items for an adaptive test of 30 items intended to be parallel to the existing P&P form of 60 items should be reasonable.

As for the large pool of 720 items, the depth of the pool was easily justified in theory; according to Green (1983) and Kingsbury (1997) the bigger the better.

III. Desired Maximum Exposure Rates

Research has shown that it is almost always problematic if an item exposure rate is more than .30 in that at least one out of three test-takers will see the same item, making test security a great concern. For most operational adaptive testing programs, an exposure rate of .20 is usually the desired limit. For example, the desired maximum exposure rates were all set at .20 in the case studies of the CAT versions of the GRE General (Eignor, Stocking, Way, & Steffen, 1993) and the comparability study of the GRE General CATs (Schaeffer et al., 1995).

Two target maximum desired exposure rates of .10 and .20 were studied. Except for the 5-4-3-2-1 algorithm where there is no direct control over the item exposure, the four exposure control algorithms described previously maintained the maximum observed appearance rates for most used items in the pool below .10 and .20 respectively, and these maximums were controlled throughout the pool in this study.

Specification of Decisions for the CAT Components

The decisions for some CAT components have been conventionally adopted in accordance with common practice, but vexing arguments still center around the choices for some other components. It is also inevitable that a decision for one component will interact with decisions for others. The design of the current study was based upon an appropriate combination of the choices for the CAT components. The decisions were made with respect to the plausibility and popularity of the options, as well as in accordance with the purposes of the study. It was anticipated that the design of the current simulation study was fairly typical of the preparation for large-scale CAT implementations, so the results could be generalized to real testing situations in general.

The option for each of the procedural CAT components is specified below.

I. The Item Response Model and the Item Pool

Real item pools were employed which contained the same item parameters as the existing P&P ACT-Math test calibrated by using the three-parameter logistic (3-PL) model (Birnbaum, 1968) and the implementation of BILOG Version 3 (Mislevy & Bock, 1990) on a single ability scale. Random samples of a broad range of 2000+ examinees responding to each item were used to estimate the item parameters for different forms of the ACT-Math test. The obtained estimates of item parameters were assumed to be the true parameters of the items.

II. The Starting Point

The fixed starting point rule was adopted for the present study to simulate a situation where no a priori information is available about the individuals. Each simulee was assigned an ability estimate of 0 to initiate the testing process.

III. The Ability Estimation Method

Among the ability estimation methods that are commonly encountered in the CAT literature, maximum likelihood estimation (MLE) (Birnbaum, 1968) is deemed appropriate for estimating examinees' final abilities, mainly because MLE produces measures that are the most unbiased (Hsu & Tseng, 1995, Wang, 1995; Wang, 1997). Accordingly, this method was adopted in the current study. Owen's Bayesian strategy (Owen, 1975) was employed for the provisional ability estimation because MLE arbitrarily estimates the ability for the all-correct or all-incorrect response patterns.

IV. The Item Selection Procedure

Because it is theoretically appealing and has wide applications, the maximum information item selection (Lord, 1977, 1980) criterion was employed in this study. In order to apply the current simulation results to a realistic testing environment, this study took into account content

constraints during the item selection process. The content balancing issue was handled by incorporating the Kingsbury and Zara mechanism (1989) into the maximum information item selection procedure. The first item was administered following the rules of item exposure control, regardless of the item's content attribute. The percentage of items that had been administered in each content area was computed and compared to the corresponding pre-specified percentage. The domain with the largest discrepancy between the empirical and the desired percentages was then identified, from which the next item was selected and administered based on the various exposure control algorithms.

In the current study, the desired content coverage of the CATs was established in accordance with the percentage of the conventional ACT-Math test items for each content attribute. The ACT-Math tests are developed according to the content specifications on these six areas: pre-algebra (23%), elementary algebra (17%), intermediate algebra (15%), coordinate geometry (15%), plane geometry (23%), and trigonometry (7%).

V. The Test Termination Rule

The fixed test length rule was adopted to terminate an adaptive testing session. On one hand, the content presented to each examinee was more directly controlled with a fixed-length adaptive test. On the other hand, the variable-length CATs could result in unacceptably large biases in the final ability estimates (Stocking, 1987). The length of the test was fixed at 30 items.

Procedures for Data Simulation

For each combination of the exposure control methods, the pool sizes, and the desired maximum exposure rates, a sequence of simulations was performed. Within each sequence of simulations, two stages were involved. First, the exposure control parameter for each item in the pool was developed according to the specific exposure control algorithm and the CAT design proposed for this study. This stage did not apply to the 5-4-3-2-1 method nor the method of exercising no exposure control due to no utilization of the parameters. Then, adaptive tests were administered to simulees in the operational testing situations. The same CAT design was followed.

I. The Stage of Developing Exposure Control Parameters

To determine the exposure control parameters for the items, this stage proceeded in the following steps.

Step 1. For the Sympson and Hetter procedure, the Davey and Parshall methodology, and the Stocking and Lewis unconditional procedure, the adaptive tests were administered to a sample of 50,000 examinees drawn from a normal distribution with a mean of 0 and a variance of 2 (i.e., $N(0,2)$) on the θ metric. The huge sample was used to make sampling error of little concern. A normal distribution is, in many cases, a representative ability distribution of real examinee

populations. The reason for using a $N(0,2)$ instead of a standard normal distribution, $N(0,1)$, was to allow more examinees at the two ends of a typical ability range of interest in the development of exposure control parameters. By using an ability distribution of greater variability, the exposure control parameters of items especially informative at the extreme ability levels were developed based on more simulees and the effect of sampling error could be mitigated (T. Davey, personal communication, October 15, 1997).

For the Stocking and Lewis conditional multinomial procedure, the development of exposure control parameters was in reference to a particular level of proficiency. The adaptive tests were administered to a conditional sample of 3,000 examinees at each of the θ levels equally spaced over the interval of -3.2 and 3.2 with an increment of .40 (i.e., -3.2, -2.8,..., 3.2), totalling 17 ability points. This can be regarded as administering the adaptive tests to a sample of 51,000 examinees (with 3,000 examinees at each of the 17 ability levels) whose abilities were uniformly distributed over the range of -3.2 and 3.2 on the θ scale.

Step 2. During the process of administering adaptive tests to the sample of 50,000 examinees or to the conditional sample of 3,000 examinees at each θ point, tentative item exposure parameters for the various exposure control techniques were developed according to the specific algorithms. However, for the unconditional and conditional multinomial procedures of Stocking and Lewis, the desirability of items was ordered only based on item information values, not on the weights as specified in the original procedures for which the Stocking/Swanson WDM (Stocking & Swanson, 1993; Swanson & Stocking, 1993) was employed to select items. The content was balanced by choosing items from adequate domains, following Kingsbury and Zara's mechanism (1989). After the whole sample had been tested, one round of iteration was completed.

Step 3. The process repeated step 2 by administering adaptive tests to the same sample of examinees, and the exposure control parameters for the various strategies were updated. The iterations continued until the observed maximum exposure rates were approximately equal to the desired level and the exposure control parameters were stabilized in the subsequent iterations. The stabilized parameters at the final round of iterations were the exposure control parameters to be used in real adaptive testing.

II. The Stage of Administering CATs in Operational Testing Situations

At this stage of administering CATs in operational testing situations, the simulation proceeded in one cycle. Each of the five exposure control techniques was incorporated into the item selection procedure under the conditions of different pool sizes and different desired maximum exposure rates. Also, the item selection with no exposure control followed the steps as well.

Step 1. A sample of 50,000 simulees drawn from a standard normal distribution, representing the real examinee population, was administered the tests.

Step 2. The adaptive test was delivered to each examinee of the representative sample, following the CAT design established for this study. Items were selected and administered according to the specific algorithm of an exposure control strategy. For those methods that directly control item exposure, the exposure control parameters developed from the previous stage were utilized here to manage the administration frequencies of the selected items. The content presented to the examinees was balanced according to Kingsbury and Zara's algorithm (1989).

Step 3. The adaptive tests were also administered to a conditional sample of 3,000 examinees at each of the ability points equally spaced over the range of interest between -3.2 and 3.2, at intervals of size .40. This step was carried out in order to obtain the conditional maximum observed exposure rates and the test-retest overlap rates (to be defined in the discussion of the methods for data analyses), as well as to evaluate the measurement precision conditional on each θ point.

Methods for Data Analyses

Three criteria were used to evaluate the strengths and weaknesses of the exposure control procedures under the various conditions specified previously. Their effectiveness as exposure control tactics were determined by item security, item overlap and CSEM of the ability estimates. The results with no exposure control exercised are presented and compared to those under control with respect to each of the criteria.

I. Item Security

As defined in Stocking and Lewis (1995a), item security is indicated by the maximum probability of administration observed for any element in the item pool. A single number summary of item security in reference to the entire examinee group was reported for each of the exposure control methods. The averages and standard deviations (SD) of the observed exposure rates over all items that were used at least once were compared.

In addition, the observed maximum probability of administration with respect to examinees of a particular ability level was reported to show the degree of test security achieved by each method conditionally.

II. Item Overlap

As specified by Davey and Parshall (1995), the test overlap rate can be classified into the test-retest overlap rate and the peer-to-peer overlap rate. These values show the extent to which

pairs of items appear together across tests taken by examinees of the same ability or differing abilities.

The test-retest mean overlap rate was obtained by first computing the percent of items that overlap between adaptive tests given to an examinee of ability θ twice, then averaging the overlap percentages over all examinees at this θ level. The peer-to-peer mean overlap rate was obtained by first calculating the overlap percentage of tests taken by two examinees generated from a $N(0,1)$ distribution, then averaging the overlap percentages over all paired examinees.

III. Conditional Standard Error of Measurement (CSEM)

At each θ point, the standard error of measurement (SEM) was calculated by the formula:

$$SEM(\theta) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta} - \bar{\theta}\right)^2},$$

where $r = 1, 2, \ldots, R$ is the number of replications (i.e., the number of adaptive tests administered) at θ, and $\bar{\theta} = \frac{1}{R}\sum_{r=1}^{R}\hat{\theta}$ is the mean of the ability estimates over the R replications at θ.

RESULTS AND DISCUSSION

Description of the Item Pools

Table 1 shows the summary statistics of the item parameters for the small and the large pools. For the small pool, the overall mean and SD of the a parameters were 1.02 and .33; the mean and SD of the bs were .19 and 1.04; and the mean and SD of the cs were .18 and .08. The large pool consisted of items of very similar quality--the mean and SD of the a parameters were 1.02 and .33; the mean and SD of the bs were .16 and 1.06; and the mean and SD of the cs were .17 and .08.

In both item pools, there existed more items at the middle of the ability continuum than at the two ends. Also, both pools contained items that discriminate moderately well with more discriminating power at the middle difficulty levels than at the extreme levels. The inspection of these two pools reveals the nature of item parameter characteristics in practical settings. Based on these two real item pools, the generalization of the simulation results in this study was more significant to operational situations than if ideal item pools had been utilized.

The Results of Developing Exposure Control Parameters

The exposure control parameters were developed through a series of adjustment simulations for the Sympson and Hetter procedure, the Davey and Parshall methodology, and both

the Stocking and Lewis unconditional and conditional multinomial procedures. Displayed in Figures 1 through 4 are the observed maximum exposure rates at each iteration stage for the four procedures under the two item pool sizes and the two desired exposure rates. The horizontal line is the desired exposure rate of $r = .10$ or .20. The observed maximum exposure rates are presented to illustrate the converging status of the exposure control parameters.

The numbers of iterations carried out for these four procedures in developing the exposure parameters were not the same. The decision for the various iterative steps that were deemed necessary and appropriate for being able to detect whether the exposure control parameters converged was based on the suggestions made in the original studies of the respective algorithms and some preliminary tryouts employing the CAT procedures designed for the current study.

Described below are the results of the iterations to develop the exposure control parameters for the Sympson and Hetter, the Davey and Parshall, and both the Stocking and Lewis unconditional and conditional multinomial procedures. The amount of time approximately needed for one round of iteration for the various procedures is reported based on the CAT design of the current study, as well as on the computer power of 233MHz (AMD-K6 MMX Enhanced Processor, 512 KB Cache) with 48MB of RAM.

The Sympson and Hetter Procedure

The development of the exposure control parameters for the Sympson and Hetter method was performed through 25 iterations. It took about 5.3 minutes to finish one iteration for the small pool and about 7.7 minutes for the large pool.

Figures 1 through 4 contain the iteration results for the Sympson and Hetter procedure under the four conditions. The maximum observed probabilities converged smoothly to the pre-specified value of .10 or .20. For the desired exposure rate of .10, the observed maximums were stabilized after the 12th iteration for the small pool and after the eighth iteration for the large pool. When the desired exposure rate was specified as .20, the iterations converged several steps faster--after six iterations for the small pool and five iterations for the large pool.

The Davey and Parshall Procedure

Davey and Parshall's iterative steps were repeated 120 times. On average, each iteration took about 6.7 minutes under the small pool condition and about 8.3 minutes under the large pool condition. Compared to the Sympson and Hetter iterations, the Davey and Parshall iteration simulations were much more time-consuming, not only because more time was needed in one iteration, but especially because many more iterative steps were required for this procedure to converge.

As shown in Figures 1 through 4, the observed maximum exposure rates converged to a value very close to the pre-specified rate of .10 or .20. For the target maximum probability of .10, the rates became stabilized after 60 iterations for the small pool and 45 iterations for the large pool. For the target probability of .20, the rates were stabilized after 30 iterations for both pools.

The Stocking and Lewis Unconditional Multinomial Procedure

The adjustment simulations for the Stocking and Lewis unconditional multinomial procedure were conducted through 25 iterations. For one iteration, approximately 22 minutes were needed for the small pool and 43.8 minutes were needed for the large pool. The average amount of time consumed for one iteration of this procedure was much greater than that for the above two procedures.

Figures 1 to 4 also display the results of the adjustment simulations for this unconditional multinomial procedure . It can be noticed that the iteration curves of this procedure were almost identical to those of the Sympson and Hetter procedure under all conditions. The observed maximum exposure rates converged at almost the same stage as in the Sympson and Hetter iterations. Because these two methods employed the same algorithm of adjusting exposure control parameters, with the only difference being in the way an item was selected, these findings reveal that the two ways of item selection would not lead to differences in the outcomes of the iteration simulations.

The Stocking and Lewis Conditional Multinomial Procedure

Thirty iterations were carried out for this procedure. More iterative steps than those of its unconditional counterpart were needed mainly because the conditional exposure parameters were derived in reference to a particular proficiency level instead of the whole examinee group. Fewer items appropriate for administration for both extreme ability levels would hinder the converging process for this conditional strategy. The amount of time needed for one iteration was approximately 19 minutes for the small pool and about 42 minutes for the large pool. The entire derivation process was very time-consuming[1].

The observed maximum exposure rates for the iterative simulations are plotted separately for each of the 17 ability levels in Figures 1 to 4. It can be noticed at a glance that these iteration results were very different from those of the other procedures. First, the values did not approach the pre-specified values as closely as those of the above three procedures. Second, the conditional

[1]According to Stocking and Lewis (1995b), a substantial amount of time could be saved if the initial exposure control parameter for each item is set to a value close or equal to the desired rate. Their suggestion was not adopted for the current study to enable comparisons among methods.

observed maximum exposure rates for the various ability levels seem to converge to different values. A careful examination of the figures for the various ability levels reveals that not only were the conditional observed maximums for the extreme ability levels further away from the pre-specified value than those for the middle ability levels, but they also appeared to converge less smoothly. Third, the iteration configurations show that this conditional multinomial procedure took a few more iterative adjustment simulations for the maximum observed exposure rates to stabilize than the Stocking and Lewis unconditional procedure or the Sympson and Hetter approach. These results were intuitively not very favorable, but they reflected the nature of the real item pools, in that there were more items appropriate for the middle ability levels than for the two extremes.

One problem was detected when the Stocking and Lewis conditional multinomial procedure was applied to the small item pool and the target exposure rate of .10. Approximately 1% of the simulees received incomplete adaptive tests; that is, their adaptive tests were terminated before the full tests of 30 items were presented. These results indicate that a full length adaptive test cannot always be guaranteed by the multinomial selection model where the ordered list consists of all items in the pool. The solution provided by M. L. Stocking (personal communication, February 23, 1998) to remedy this situation is based on the idea that the next item is selected from a list of items that does not include all available items. Assuming that all items are discrete for simplicity, the length of this list is determined by dividing the item pool size by the adaptive test length. The integer value of this ratio is the length of the list from which the next item is selected. Based on Stocking's experience, this rule provides satisfactory results of complete test administration.

The Results of Administering CATs in Real Testing

The exposure control parameters that resulted from the final round of the iterations were associated with the corresponding items to be used in the operational testing situations. In addition to the above four methods for which the exposure control parameters are utilized to directly limit the overexposure of items, the 5-4-3-2-1 randomization technique of McBride and Martin and the procedure with no control over item exposure were also incorporated into this stage using the same CAT design.

While the no control procedure and the randomization technique are included in the tables or figures under the desired rate label of .10 or .20, they only serve to facilitate comparisons with the other procedures under these two desired maximum conditions. The maximum desired rates of .10 and .20 did not affect the performance of these two procedures due to no utilization of the exposure control parameters. The differences between the results reported under the desired rates of .10 and .20 were simply due to sampling errors.

Described below are the results of all six procedures under the various conditions with respect to each of these criteria: the item security, the test-retest and the peer-to-peer test overlap rates, and the CSEM. The analyses for the overall test security and the peer-to-peer test overlap rates were performed by administering the adaptive tests to examinees drawn from a N(0,1) distribution. To compute the indices of the conditional observed maximums, the test-retest overlap rates and the SEMs conditional on each ability level, the adaptive tests were administered to conditional sample sizes of 3,000 examinees over the 17 equally spaced ability points from -3.2 to 3.2. The evaluation process was based on general observations of the results, and the results of using the small pool vs. the large pool, as well as the results of specifying the desired maximum exposure rate as .10 or .20.

I. Item Security

Table 2 reports the summary statistics of the observed exposure rates as a result of applying the various exposure control methods to the entire examinee group. The N column lists the numbers of items that were administered to examinees at least once. Based on these items, the summary statistics were obtained. The column of the maximum exposure rate shows the degree of item security achieved by the various procedures in light of the examinee group as a whole.

In addition, the results were also analyzed in a conditional fashion. The maximum exposure rates conditionally observed at each ability level are plotted to show test security at the respective ability levels.

General Observations

The effect of the difference in the examinee populations for the derivation of the exposure control parameters and the operational real testing was directly reflected in the maximum exposure rates (i.e., item security indices) observed for the Sympson and Hetter, the Davey and Parshall, and the Stocking and Lewis unconditional procedures. In order to allow more simulees at the two ends of a typical ability range of interest in developing the exposure control parameters, the iterations were carried out by administering the adaptive tests to examinees from a normal distribution with a mean of 0 and a variance of 2.

However, the final exposure control parameters were applied to examinees coming from a standard normal distribution in operational testing situation. The difference of the examinee populations caused the observed maximums resulting from using any of the above exposure control methods to be higher than the expected rates, although the exposure control parameters for any of these methods converged and the observed maximum exposure rates approached the pre-specified rates in the derivation stage.

The advantage of employing the multinomial procedure conditional on ability level was obvious, as seen in Table 2. The conditional exposure parameters were derived in reference to a particular ability point, so the exposure rates of items were directly controlled at the respective levels of examinee ability. This procedure was, therefore, independent of the target examinee population in real testing. The maximum rates were observed as low as the target exposure rates of .10 or .20.

The conditional maximum observed exposure rates at each ability level are shown in Figure 5. As displayed in these plots, while controlling the item exposure rate unconditionally with randomization, the Sympson and Hetter, and the Stocking and Lewis unconditional procedures, the conditional observed maximums reached as high as 1.0, particularly at the two extreme ability levels. This high degree of insecurity was mitigated at the middle range of the ability continuum, but the maximums were still around .30 and .40 for the target exposure rate of .10. The condition deteriorated when the target probability of administration was set to be .20; the conditional observed maximums were all greater than .60.

The conditional observed maximums for the Davey and Parshall method appeared to be fairly stable over most of the ability continuum, although they reached .30 and .40 for the desired rates of .10 and .20 respectively. The results of employing the Stocking and Lewis conditional procedure were the most satisfactory. The values were slightly above the desired levels for the middle ability points, which accounted for most part of the examinee population. The maximums were observed somewhat higher at both extremes, as expected to be consistent with the results of the final round of the Stocking and Lewis conditional iterations.

The Small Pool vs. The Large Pool

As shown in Table 2, while the item selection was not controlled for exposure rates, the maximum observed rate reached 1.0 for both small and large item pools. This finding suggests that a large pool itself was not sufficient to guarantee test security. When the selection was based on the randomization scheme, the employment of the large pool only led the maximum observed to drop from .72 to .64. That is, even with the large pool of 720 items, the randomization procedure still resulted in more than six out of ten examinees being exposed to the same item.

As mentioned previously, the slightly higher values of the maximum observed rates than desired were due to the difference in the examinee populations. However, it is somewhat counter-intuitive that as the pool size was doubled, the maximum observed exposure rates for the various methods, except for the Stocking and Lewis conditional multinomial procedure, were higher than those for the small pool. A further examination of the behavior of the items under the various exposure control algorithms reveals that items that were the most frequently administered in the small pool were no longer administered as often in the large pool, and items of the most often

administered in the large pool were either those not contained in the small pool or those with high observed exposure rates in the small pool. These findings suggest that there existed a few items in the large pool that were especially appropriate for administration. Subject to the employment of the particular exposure control algorithm, these items might be administered relatively often to cause the higher appearance frequencies in the large pool. Or, the higher observed maximums in the large pool than in the small pool might be simply due to sampling errors.

As the pool size was doubled, the average exposure rate decreased for each method (see Table 2). These results were expected since there were twice as many items of similar quality in the large pool to allow more choices for item selection for administration. Based on the way the average exposure rates were calculated in this study, the average exposure rate would be lower as more items were administered at least once. For the Davey and Parshall and the Stocking and Lewis conditional procedures, the average observed exposure rates were reduced by half in the large pool (.08 to .04 for the desired rate of .10 and around .08 to around .05 for the desired rate of .20). The pool utilization in the large pool for the Sympson and Hetter and the Stocking and Lewis unconditional procedures did not increase as much as that for the Davey and Parshall and the Stocking and Lewis conditional procedures. The average values were only reduced from .09 to .07 for the desired rate of .10 and from .12 to .11 for the desired rate of .20. It seems that the Davey and Parshall and the Stocking and Lewis conditional procedures were more likely to benefit from enlarging the pool size than the other methods.

Investigating the maximum exposure rates conditionally observed reveals a similar phenomenon to that observed in the overall maximums. As displayed in Figure 5, the conditional maximum values were still observed to be higher with the large pool than the small pool for all or most ability levels, indicating that the utilization of the large pool did not profit the procedures in achieving higher test security. Again, except for the sampling errors, these results might be caused by having some items in the large pool particularly appropriate for administration for some ability levels.

The Desired Maximum Exposure Rate of .10 vs. .20

For the target probability of .10, the Sympson and Hetter, the Davey and Parshall and the Stocking and Lewis unconditional procedures yielded very similar maximum observed exposure rates, which were all slightly greater than the pre-specified rate (see Table 2). The Stocking and Lewis conditional procedure successfully controlled the item exposure at the desired value of .10.

For the target probability of .20, the Davey and Parshall approach controlled the maximum observed exposure to a value slightly lower than those of the Sympson and Hetter and the Stocking and Lewis unconditional procedures. Again, the Stocking and Lewis conditional procedure

successfully limited the exposure rates of items to the desired value of .20, providing evidence that this procedure was independent of the target population distribution.

Figure 5 conveys the information that as the desired exposure rate was relaxed to .20, all the procedures resulted in the conditional observed maximums greater than .40 at almost all ability levels, with the exception of the Stocking and Lewis conditional method.

II. Item Overlap

The performance of each method in terms of the test-retest overlap rates and the peer-to-peer overlap rates is evaluated below.

1. The Test-Retest Overlap Rate

The test-retest overlap rates were expected to be high at each ability level since these values were computed based on examinees of the same ability retaking the tests without intervention and items administered to them were likely to be selected, for the most part, from the same part of the item pool. The test-retest mean overlap rates for each of the 17 ability levels are shown in Figure 6. The overall test-retest mean overlap rates reported in Table 3 are the average values of the test-retest mean overlap rates at all ability levels. Figure 7 conveys the full distributions of the test-retest overlap rates across 17 ability levels.

General Observations

For each method, the overall patterns of the test-retest mean overlap rates were similar across the four conditions (see Figure 6). The greatest difference of the test-retest mean overlap rates between the no control and the randomization techniques was observed at the very high ability levels. The differences became smaller towards the middle ability levels. The Sympson and Hetter and the Stocking and Lewis unconditional procedures yielded identical test-retest mean overlap rates at all ability levels, with smaller values within the ability range of -1.2 and .40 and larger values at the two ends, especially higher at the very high end. The Davey and Parshall and the Stocking and Lewis conditional methods produced substantially smaller test-retest mean overlap rates, which were fairly stable across the entire ability continuum.

Figure 7 displays the entire distributions of the test-retest overlap rates for each method. For the no control and the randomization techniques, the distributions were negatively skewed for both item pools, suggesting a large proportion of high test-retest overlap values. For the Sympson and Hetter and the Stocking and Lewis unconditional methods, as the exposure rates were tightly controlled at the pre-specified rate .10, the test-retest overlap rates were low to moderate (around .10 to .50) for the most part. However, as the expected exposure rate was loosened to .20, moderate to high values of the test-retest overlap rates (around .40 to .80) were seen across most

of the distribution. The Davey and Parshall and the Stocking and Lewis conditional procedures successfully controlled most test-retest overlap values to be around .10 and .20.

The Small Pool vs. The Large Pool

As can be seen in Table 3, the overall test-retest mean overlap rates decreased slightly as the pool size was doubled for the no control, the randomization, the Davey and Parshall, and the Stocking and Lewis conditional methods. On the contrary, it was confusing to see that the large pool caused the test-retest mean overlap rates for the Sympson and Hetter and the Stocking and Lewis unconditional strategies to increase. For the desired rate of .10, the overall values for these two methods were increased from .28 to .34 as the large pool was employed. For the desired rate of .20, there resulted a fairly small difference between the overall value of .53 for the small pool and .55 for the large pool; however, these values were fairly high.

One possible explanation for this phenomenon that occurred in the Sympson and Hetter and the Stocking and Lewis unconditional methods may be similar to that observed in Table 2 where the maximum exposure rates were higher under the large pool condition than the small pool. There might exist in the large pool some items especially appropriate for examinees at some ability levels to cause the higher appearance frequencies, although still under the control of the exposure control parameters. It might be the case that the specific nature of these two algorithms was subject to the existence of such items in the large pool so that higher overlap rates were yielded.

The Desired Maximum Exposure Rate of .10 vs. .20

The configurations in Figure 6 display that for the Sympson and Hetter and the Stocking and Lewis unconditional procedures, relaxing the desired maximum exposure rate to .20 caused the test-retest mean overlap rates at each ability level to increase by a large amount. Table 3 shows that their overall test-retest mean overlap rates were increased from .28 to .53 for the small pool and from .34 to .55 for the large pool. For the Davey and Parshall method, relaxing the desired rate from .10 to .20 caused the overall test-retest mean overlap rates to increase from .16 to .19 for the small pool and from .15 to .17 for the large pool. Apparently, these increases were significantly smaller than those for the Sympson and Hetter as well as for the Stocking and Lewis unconditional strategies.

For the Stocking and Lewis conditional method, the penalty of relaxing the desired exposure rate to .20 on the increase of the test-retest mean overlap rate was very consistent across all ability levels. For either of the pools, the overall test-retest mean overlap rate was increased from .10 to .19 due to the higher exposure rate allowed. The Stocking and Lewis conditional procedure resulted in the most satisfactory test-retest mean overlap rates, followed by the Davey and Parshall methodology.

2. The Peer-to-Peer Overlap Rate

The values of the peer-to-peer overlap rates were expected to be smaller than the test-retest overlap rates, because these values represent the overlap percentages of items administered to examinees in the group as a whole. As shown in Table 4, the mean peer-to-peer test overlap rates for the various methods were substantially smaller than the overall test-retest mean overlap rates reported in Table 3.

In addition to the descriptive statistics of the peer-to-peer overlap rates summarized in Table 4, their full distributions are displayed in Figure 8 for each method respectively.

General Observations

As presented in Table 4, as long as the selection procedure did not control for item exposure or the control was solely based on the randomization scheme, the peer-to-peer overlap rates could reach a value as high as 1.0, no matter whether the small or the large pool was used. On average, the peer-to-peer overlap rates for both the no control and the randomization techniques were about .37 for the small pool and .34 for the large pool, indicating that the adaptive tests for any two examinees of this group contained more than ten identical items. The employment of the Sympson and Hetter and the Stocking and Lewis unconditional procedures could also yield very high values of the peer-to-peer overlap rates. The maximum peer-to-peer overlap rates were relatively lower for the Davey and Parshall and the Stocking and Lewis conditional procedures.

The full distributions of the peer-to-peer overlap rates in Figure 8 show that for the no control and the randomization procedures, these types of overlap values were evenly distributed over almost the entire range. The distributions for the Sympson and Hetter and the Stocking and Lewis unconditional approaches were identical. For the Davey and Parshall and the Stocking and Lewis conditional procedures, small overlap rates of .10 to .20 were observed for the majority of the distribution.

The Small Pool vs. The Large Pool

The employment of the large pool caused the average peer-to-peer overlap percentages to be reduced by about 3% for the no control and the randomization techniques. For either of the two desired exposure rates, the mean peer-to-peer overlap rates for the Sympson and Hetter and the Stocking and Lewis unconditional methods were similar under the two different pool sizes. The employment of the large pool lowered the average overlap rates for the Davey and Parshall method from .10 to .09 when the expected rate of .10 was specified, and from .14 to .11 when the expected rate of .20 was used. For the Stocking and Lewis conditional method, using the large pool also led to a reduction in the mean overlap rates--from .09 to .06 for the expected rate of .10 and from .12 to .09 for the expected rate of .20.

The Desired Maximum Exposure Rate of .10 vs. .20

The relaxation on the desired maximum rate from .10 to .20 approximately doubled the average peer-to-peer overlap rate for both the Sympson and Hetter and the Stocking and Lewis unconditional procedures. Such a big penalty was not seen in the Davey and Parshall or the Stocking and Lewis conditional methods. For the small pool, loosening the desired rate increased the overlap rates from .10 to .14 for the Davey and Parshall method and from .09 to .12 for the Stocking and Lewis conditional procedure. For the large pool, the overlap rates were increased from .09 to .11 for the Davey and Parshall method and from .06 to .09 for the Stocking and Lewis conditional procedure due to the higher exposure rate allowed.

III. Conditional Standard Errors of Measurement

Due to the constraints on test content, measurement precision for the current CATs was expected to be compromised. Although the measurement precision might have been sacrificed to some extent, examinees at any level of ability were ensured to receive CATs with appropriate content balance.

General Observations

The CSEM curves produced by using the various methods under the different pool sizes and the different expected exposure rates are displayed in Figure 9. The large conditional sample sizes of 3,000 examinees led to the smooth curves in the configurations. Because there were fewer items appropriate for administration for both extreme ability levels, the SEMs at the two ends were higher than those at the middle part of the ability scale. The effect of guessing caused higher SEMs at the lower end than at the upper end of the scale.

Figure 9 displays that, across all conditions, the Davey and Parshall and the Stocking and Lewis conditional strategies resulted in higher CSEMs than the no control, the randomization, the Sympson and Hetter and the Stocking and Lewis unconditional procedures. These results were stronger at the two extremes than at the middle range. The price for better control of the item exposure rates with the Davey and Parshall and the Stocking and Lewis conditional algorithms than with the other approaches was seen in the loss of measurement precision. Increases in the CSEMs for these two methods were not unexpected.

Similar to the plots displayed for the previous criteria, the CSEM curves for the Sympson and Hetter and the Stocking and Lewis unconditional procedures were almost identical at any ability point, supporting the argument that these two procedures were not different from each other. Also, the randomization technique resulted in very similar CSEMs to those of the selection without exposure control, conveying the information again that employing the random selection from among a group of items was not different from the no control procedure.

The Small Pool vs. The Large Pool

Investigating the CSEMs between the small pool and the large pool in Figure 9 reveals with no surprise that such random errors decreased for all methods as the large pool was employed. However, the degree to which the CSEMs decreased was not very substantial if the control for item exposure was not exercised at all or was simply based on the randomization scheme. The reduction in the CSEMs seems more noticeable for the four procedures which directly incorporated statistical methods into the item selection process, particularly with the desired rate of .10.

A careful inspection of Figure 9 suggests that the Davey and Parshall and the Stocking and Lewis conditional methods seem more likely to take advantage of enlarging the pool size to remedy the loss of measurement precision than the Sympson and Hetter and the Stocking and Lewis unconditional procedures, although the difference was very small.

The Desired Maximum Exposure Rate of .10 vs. .20

When the desired rate was specified as .10, the Stocking and Lewis conditional strategy produced the highest SEMs at any ability level for both pools. For the small pool, its CSEMs were substantially higher than those of the other methods. Such high values of CSEMs (i.e., low reliability) were especially prevalent at the two extremes of the continuum. This phenomenon could be explained by the fact that this conditional method strictly limited the exposure rates of items appropriate for administration at these extreme ability levels where, in fact, there existed not many of these kinds of items in the pool. Also, the fact that about 1% of the examinees received incomplete adaptive tests might contribute in part to these high CSEMs.

The high CSEM curves of the Stocking and Lewis conditional method for the desired rate of .10 were followed by the Davey and Parshall method for both item pools. It can be seen in Figure 9 that for the middle ability levels, the random errors for the Davey and Parshall method were very close to those for the Sympson and Hetter and the Stocking and Lewis unconditional methods. Beyond this range, the loss of measurement precision due to the employment of the Davey and Parshall procedure appeared more prevalent. However, the magnitude of the loss was judged tolerably small in reference to the degrees of test security this procedure achieved and the numbers of items this procedure utilized.

As to the Sympson and Hetter and the Stocking and Lewis unconditional procedures, while the SEMs at the middle range were as high as those of the Davey and Parshall procedure which were relatively higher than those of the no control and the randomization techniques, the SEMs at the ability points beyond -2.4 and 2.0 were as low as those of the no control and the randomization procedures.

The configurations were somewhat different as the target exposure rate was relaxed to .20. As can be observed in Figure 9, the CSEMs for the Sympson and Hetter as well as the Stocking

and Lewis unconditional procedures were very close to those of the no control or the randomization techniques. These results indicate that when the desired exposure rate was as high as .20, these two procedures would achieve the same measurement precision as the no control and the randomization procedures. On the other hand, this might be an indication that as the desired exposure rate was loosened to .20, it is likely that these two procedures would be no better in improving test security than the no control and the randomization techniques.

As expected for the Davey and Parshall and the Stocking and Lewis conditional strategies, the CSEMs were higher than those of the other four methods. In contrast to the results with the desired rate of .10 where the Davey and Parshall strategy produced lower SEMs at all ability levels than the Stocking and Lewis conditional method, there existed no difference of SEMs between these two procedures at the middle ability levels but slightly higher SEMs at the two extremes were yielded by the Davey and Parshall method instead of the Stocking and Lewis conditional procedure when the desired rate .20 was specified.

SUMMARY AND CONCLUSIONS

General Observations

The results show that the 5-4-3-2-1 randomization technique did not better ensure item security to any noticeable extent than the procedure without exposure control. Although the most informative items at a current ability estimate in the early testing process were not always administered, items administered later were not controlled for their appearance frequencies and examinees might receive many of the same items subsequently. These findings suggest that the test security issue cannot be easily remedied by randomly selecting items from among a group of items at the early stages of testing process.

The Sympson and Hetter and the Stocking and Lewis unconditional multinomial procedures produced very similar results under all conditions. The patterns of their curves were almost identical with respect to each criterion. It has been described that both procedures employed the same algorithm of adjusting exposure control parameters and the only difference resided in the way an item was selected. Based on the results, the difference in how an item was selected did not lead to differences in the performance of these two procedures in controlling the item exposure rates. However, when developing the exposure control parameters, the Sympson and Hetter procedure was much more efficient.

The results yielded by the Davey and Parshall procedure were in general favorable. This procedure controlled the frequencies of item use as well as the overlap percentages across tests to a satisfactory extent; and at the same time, the magnitude of the loss in measurement precision was judged tolerably small in light of the test security this procedure achieved. The Stocking and Lewis

conditional procedure produced the most satisfactory results in terms of the maximum observed exposure rates and the item overlap rates, but this apparently came at the price of increases in CSEMs due to the strict control of item exposure.

The results have shown the effect of population differences on the performance of the exposure control algorithms. The advantage of developing exposure control parameters conditionally on each proficiency level was seen in being independent of the target examinee population in real testing. The Stocking and Lewis conditional procedure successfully controlled the maximum observed exposure rate to the desired level in reference to the entire examinee group. Due to the nature of the real item pools in that there were fewer items appropriate for examinees at the two extreme ability levels than at the middle, it was inevitable that the conditional maximum observed exposure rates at these levels were higher than the rates pre-specified.

The Effect of Item Pool Size

As long as the selection procedure did not control for item exposure or the control was solely based on the randomization scheme, optimal items could be administered to almost every examinee under both pool size conditions. A large item pool itself was therefore not sufficient to guarantee test security. As to the other procedures, the maximum observed exposure rates were higher in the large item pool than in the small pool. These results provide no evidence that the employment of the large pool would profit the exposure control methods in reducing the maximum observed exposure rates or achieving higher test security.

It was not surprising to see that while the pool size was doubled, the average exposure rates of the items were reduced. However, the extent to which the average exposure rates decreased with the employment of the large pool varied among the methods. Enlarging the item pool size seems more profitable for the Davey and Parshall and the Stocking and Lewis conditional procedures in reducing the average exposure rates than for the other procedures.

The employment of the large pool slightly lowered the test-retest and the peer-to-peer overlap rates for the no control, the randomization, the Davey and Parshall, and the Stocking and Lewis conditional procedures. But, the impact was very little in general. The Sympson and Hetter and the Stocking and Lewis unconditional strategies did not benefit from the use of the large pool.

The extent to which the CSEMs decreased with the large pool varied among the methods. The Davey and Parshall and the Stocking and Lewis conditional algorithms appeared more likely to take advantage of increasing the pool size to remedy the loss of measurement precision than the other approaches.

These findings on the pool size effect indicate that the performance of the various methods was affected differentially by the size of the item pool. This differential effect might be attributed

to the specific features of the exposure control algorithms in combination with the structure of the item pool.

The Effect of the Desired Maximum Exposure Rate

The behavior of these strategies under the two desired maximum exposure rates appeared not to be very consistent. It implies that the features of the various exposure control strategies were affected differentially by the extent to which the item exposure rates were limited. Especially noticeable was the effect of the desired exposure rate on the results of both types of overlap percentages. The Sympson and Hetter and the Stocking and Lewis unconditional procedures seem more likely to be disadvantaged by relaxing the desired exposure rate than both the Davey and Parshall and the Stocking and Lewis conditional procedures.

Conclusions

The effectiveness and psychometric properties of the various exposure control algorithms under the two item pool sizes and the two desired exposure rates were investigated and compared with respect to the test security, the test-retest and peer-to-peer overlap rates, and the CSEMs. Each method had its advantages and disadvantages, but no one possessed all of the desired characteristics. The price for better control of the item exposure rates was paid in the increase of the CSEMs. The measurement precision was sacrificed as a consequence of the strict control of the exposure rates of optimal items.

A large item pool itself did not appear sufficient to guarantee test security. Only by incorporating statistical mechanisms in the item selection procedure can the attempt to improve test security be accomplished in the CAT environment. While the item exposure was controlled unconditionally, an item's overall exposure rate might be low for examinees across the entire ability continuum, but this item might have been administered to almost all examinees at one particular ability level. Controlling item exposure conditionally at each ability level provided satisfactory results of the test security and the overlap percentages. At the same time, because the conditional exposure control parameters were derived independently of the target examinee distribution, the performance of the exposure control parameters was not affected by the differences in the examinee populations.

However, developing the conditional exposure control parameters was very tedious and time-consuming. It is conceivable that much more time would be devoted if larger conditional sample sizes are used or if such complex adaptive testing situations are applied as the contexts having blocks of items based on common stimulus materials. When an entire item pool is retired after some time or if there is any significant change in test structures of the pool, the exposure

control parameters must be redeveloped. The time consumed in developing the exposure parameters is a very practical issue that can not be ignored.

Based on the similarity of the simulation results yielded by the Sympson and Hetter and the Stocking and Lewis unconditional procedures, it is possible that if the Sympson and Hetter exposure control parameters are derived in reference to each ability level, the performance of this approach might be competitive to that of the Stocking and Lewis conditional procedure, at least for the CAT design similar to this study. By doing so, substantial time savings could accrue in the preparation of the exposure control parameters. Further studies of the Sympson and Hetter "conditional" procedure might be beneficial to the practical issues in developing the exposure control parameters.

In conclusion, among the five exposure control algorithms investigated in this study, the Stocking and Lewis conditional procedure best served the purposes of controlling the observed exposure rates to the desired values as well as producing the lowest test overlap rates. However, the trade-off was that the measurement precision was sacrificed to some extent, particularly at both extreme ability levels.

REFERENCES

ACT. (1997). *ACT assessment technical manual*. Iowa City, IA: ACT, Inc.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading, MA: Addison-Wesley.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation* (Research Report 93-56). Princeton, NJ: Educational Testing Service.

Featherman, C. M., Subhiyah, R. G., & Hadadi, A. (1996, April). *Effects of randomesque item selection on CAT item exposure rates and proficiency estimation under the 1- and 2-PL models*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Green, B. F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 69-80). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hetter, R. D., & Sympson, J. B. (1995). *Item exposure control in the CAT-ASVAB*. Unpublished draft. San Diego, CA: Navy Personnel Research and Development Center.

Hsu, T. C., & Tseng, F. L. (1995). *Using simulation to select an adaptive testing strategy: An item bank evaluation program*. Unpublished draft, University of Pittsburgh.

Kingsbury, G. G. (1997, March). *Some questions that must be addressed to develop and maintain an item pool for use in an adaptive test*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*(1), 95-100.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing : Latent trait test theory and computerized adaptive testing* (pp. 223-236). New York, Academic Press.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, 9(4), 287-304.

Mislevy, R. J., & Bock, R. D. (1990). Item analysis and test scoring with binary logistic models. *BILOG 3*. Chicago, IL: Scientific Software, Inc.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*(350), 351-356.

Schaeffer, G. A., Steffen, M., Golub-Smith, M. L., Mills, C. N., & Durso, R. (1995). *The introduction and comparability of the computer adaptive GRE General Test* (GRE Board Professional Report 88-08aP; Research Report 95-20). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review, 36(3/4)*, 263-277.

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report 93-2). Princeton, NJ: Educational Testing Service.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (Research Report 94-5). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995a). *A new method of controlling item exposure in computerized adaptive testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Lewis, C. (1995b). *Controlling item exposure conditional on ability in computerized adaptive testing* (Research Report 95-24). Princeton, NJ: Educational Testing Service.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*(3), 277-292.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*(2), 151-166.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item-exposure rates in computerized adaptive testing.* Paper presented at the 17th annual meeting of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

Wang, T. (1995). *The precision of ability estimation methods in computerized adaptive testing.* Unpublished doctoral dissertation, University of Iowa, Iowa City.

Wang, T. (1997, March). *Essentially unbiased EAP estimates in computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Way, W. D. (1997, March). *Protecting the integrity of computerized testing item pools.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Table 1. Descriptive Statistics of the Item Parameters

The Small Pool

Item Parameters	N	Mean	SD	Skewness	Kurtosis	Minimum	Maximum
a	360	1.0177	0.3278	0.8661	5.3367	0.1963	2.6298
b	360	0.1925	1.0355	-0.2061	2.5611	-2.8903	2.5590
c	360	0.1751	0.0820	1.0144	4.4955	0.0401	0.5000

The Large Pool

Item Parameters	N	Mean	SD	Skewness	Kurtosis	Minimum	Maximum
a	720	1.0203	0.3322	0.6319	4.1073	0.1963	2.6298
b	720	0.1601	1.0632	-0.2540	2.8373	-3.4441	3.3984
c	720	0.1746	0.0794	0.9681	4.4833	0.0334	0.5000

Table 2. Results of Observed Exposure Rates for the Various Methods by Item Pool and Desired Exposure Rate

r=.10

	Method	N	Mean	SD	Minimum	Maximum
The Small Pool	No Control	153	0.19608	0.18658	0.00002	1.00000
	M & M	163	0.18405	0.18241	0.00006	0.72012
	S & H	347	0.08646	0.03878	0.00002	0.13640
	D & P	360	0.08333	0.03568	0.00234	0.12816
	S & L	345	0.08696	0.03810	0.00002	0.13150
	S & L C	360	0.08332	0.02284	0.00904	0.10342
The Large Pool	No Control	187	0.16043	0.17050	0.00002	1.00000
	M & M	194	0.15464	0.16745	0.00002	0.64194
	S & H	412	0.07282	0.04716	0.00002	0.14348
	D & P	654	0.04587	0.04253	0.00002	0.13536
	S & L	412	0.07282	0.04699	0.00002	0.13886
	S & L C	701	0.04280	0.02648	0.00002	0.10032

r=.20

		N	Mean	SD	Minimum	Maximum
The Small Pool	No Control	153	0.19608	0.18664	0.00002	1.00000
	M & M	165	0.18182	0.18244	0.00004	0.71930
	S & H	242	0.12397	0.09562	0.00002	0.27164
	D & P	360	0.08333	0.07027	0.00012	0.24682
	S & L	239	0.12552	0.09425	0.00002	0.26948
	S & L C	351	0.08547	0.05382	0.00004	0.19966
The Large Pool	No Control	186	0.16129	0.17057	0.00002	1.00000
	M & M	194	0.15464	0.16750	0.00002	0.64038
	S & H	279	0.10753	0.09486	0.00002	0.27338
	D & P	653	0.04594	0.05542	0.00002	0.25202
	S & L	278	0.10791	0.09558	0.00002	0.27374
	S & L C	520	0.05769	0.04391	0.00002	0.20062

Note. The descriptive statistics were based on items that were used at least once.

No Control = the no control procedure.

M & M = the 5-4-3-2-1 randomization technique of McBride and Martin.

S & H = the Sympson and Hetter procedure.

D & P = the Davey and Parshall methodology.

S & L = the Stocking and Lewis unconditional multinomial procedure.

S & L C = the Stocking and Lewis conditional multinomial procedure.

Table 3. The Overall Test-Retest Mean Overlap Rates for the Various Methods by Item Pool and Desired Exposure Rate

r=.10	No Control	M & M	S & H	D & P	S & L	S &L C
The Small Pool	0.79662	0.74673	0.28012	0.15646	0.28099	0.09770
The Large Pool	0.76264	0.72212	0.34281	0.14740	0.34290	0.09682

r=.20	No Control	M & M	S & H	D & P	S & L	S &L C
The Small Pool	0.79910	0.74762	0.53096	0.19271	0.53195	0.19082
The Large Pool	0.76265	0.72047	0.55093	0.16827	0.55099	0.18622

Table 4. The Peer-to-Peer Test Overlap Rates for the Various Methods by Item Pool and Desired Exposure Rate

r=.10

	Method	Mean	SD	Minimum	Maximum
The Small Pool	No Control	0.37110	0.28874	0.03333	1.00000
	M & M	0.36362	0.28446	0.00000	1.00000
	S & H	0.10359	0.08075	0.00000	0.60000
	D & P	0.09877	0.06610	0.00000	0.40000
	S & L	0.10401	0.08100	0.00000	0.63333
	S & L C	0.08947	0.05093	0.00000	0.33333
The Large Pool	No Control	0.34028	0.28360	0.03333	1.00000
	M & M	0.33629	0.27883	0.00000	1.00000
	S & H	0.10286	0.10690	0.00000	0.70000
	D & P	0.08500	0.07360	0.00000	0.46667
	S & L	0.10308	0.10733	0.00000	0.70000
	S & L C	0.05919	0.04991	0.00000	0.33333

r=.20

		Mean	SD	Minimum	Maximum
The Small Pool	No Control	0.37145	0.28805	0.03333	1.00000
	M & M	0.36511	0.28457	0.00000	1.00000
	S & H	0.19691	0.17375	0.00000	0.83333
	D & P	0.14258	0.08439	0.00000	0.56667
	S & L	0.19538	0.17348	0.00000	0.90000
	S & L C	0.11911	0.07757	0.00000	0.50000
The Large Pool	No Control	0.33975	0.28204	0.03333	1.00000
	M & M	0.33601	0.27786	0.00000	1.00000
	S & H	0.19200	0.19328	0.00000	0.86667
	D & P	0.11242	0.08486	0.00000	0.50000
	S & L	0.19263	0.19346	0.00000	0.90000
	S & L C	0.09006	0.07946	0.00000	0.43333

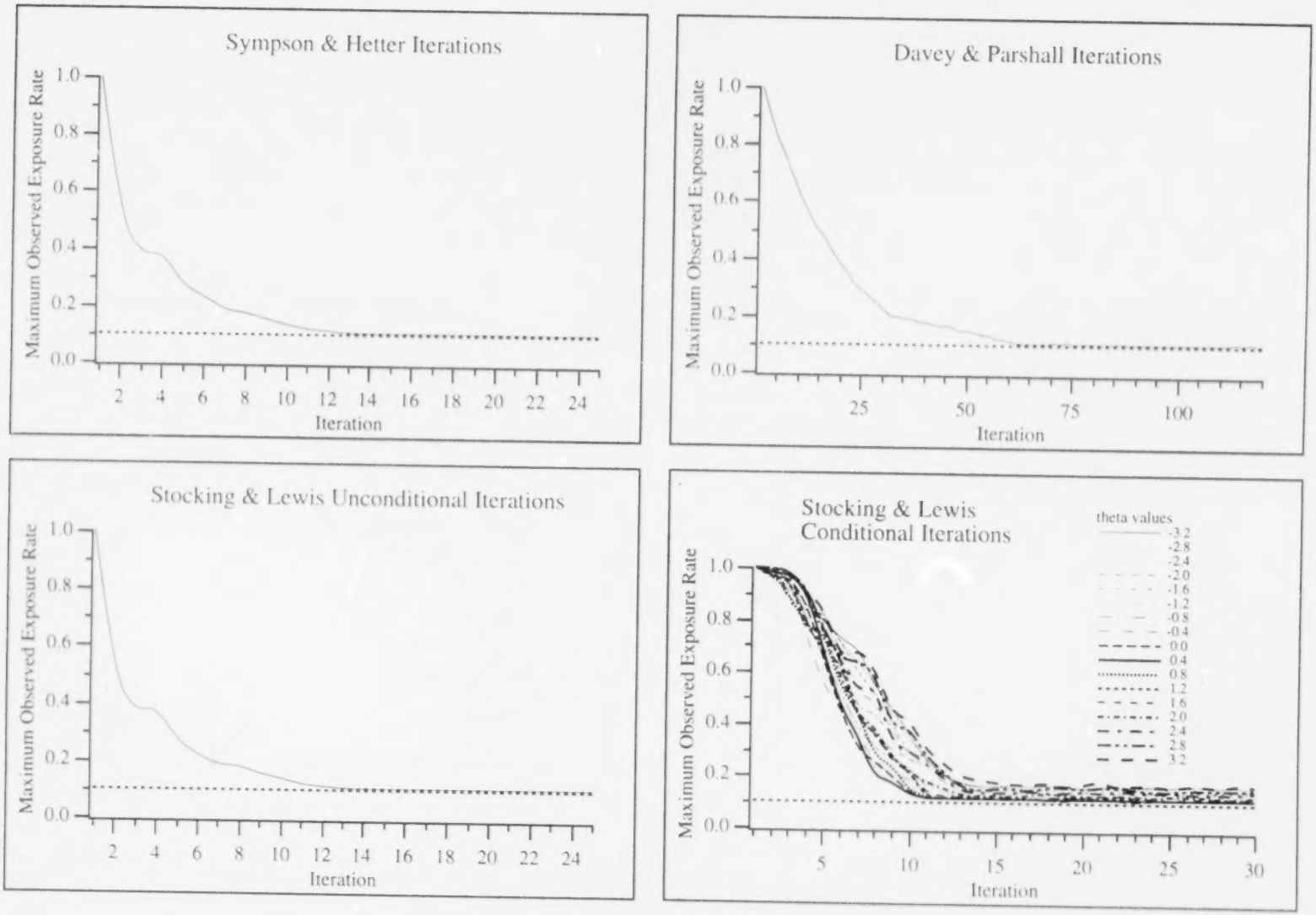

Figure 1. Results of Iterations for the Small Pool with r=.10

35

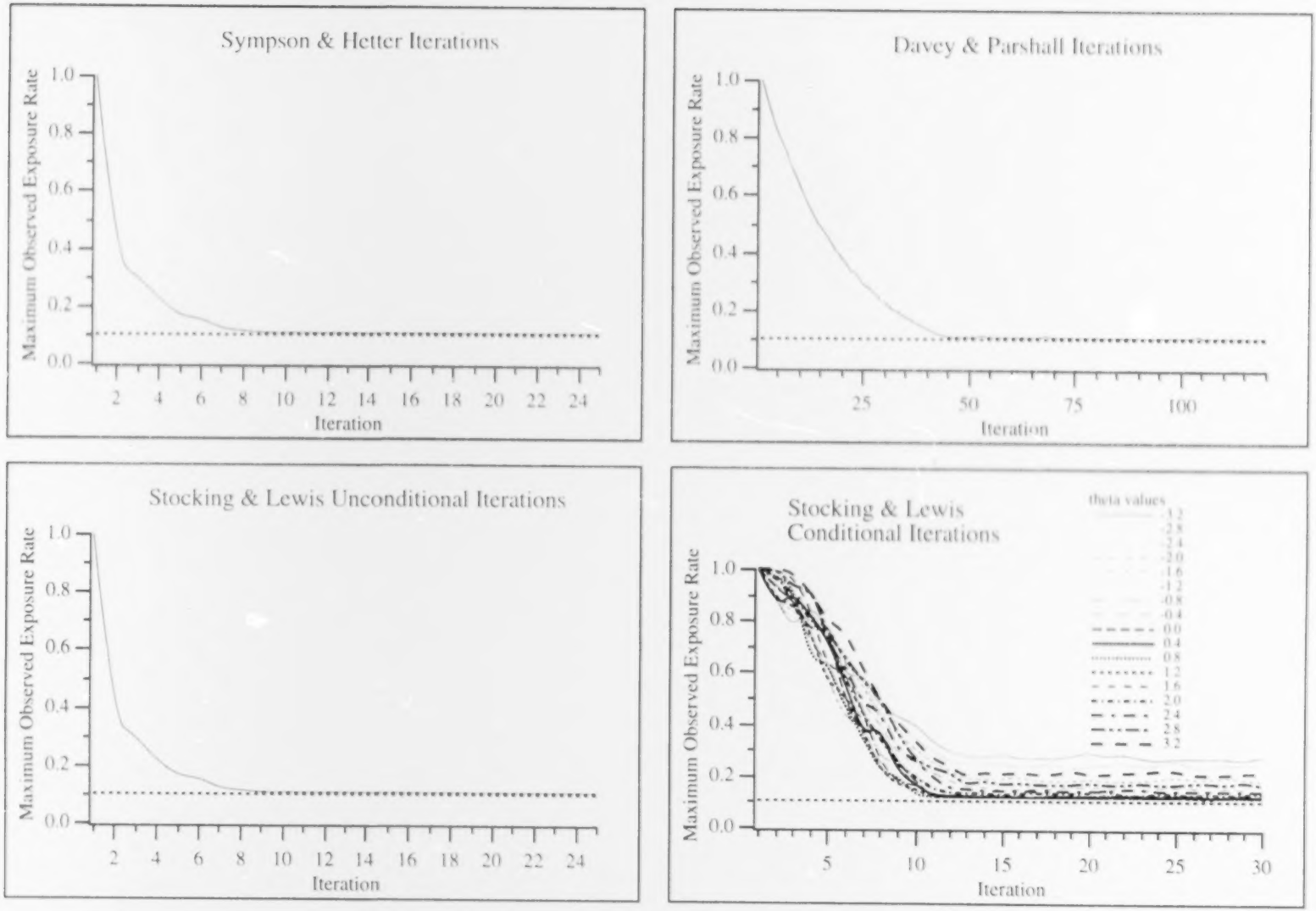

Figure 2. Results of Iterations for the Large Pool with r=.10

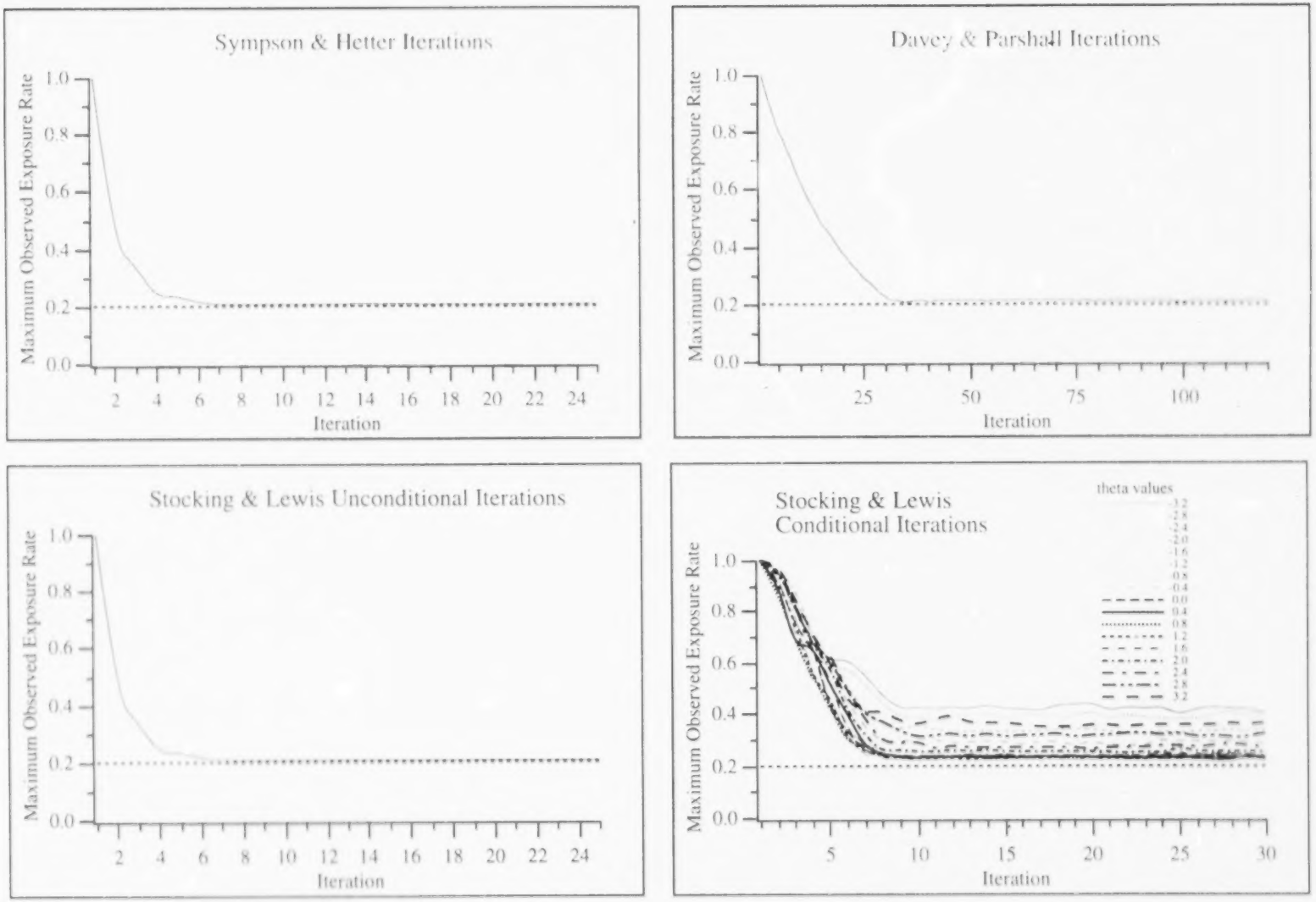

Figure 3. Results of Iterations for the Small Pool with r=.20

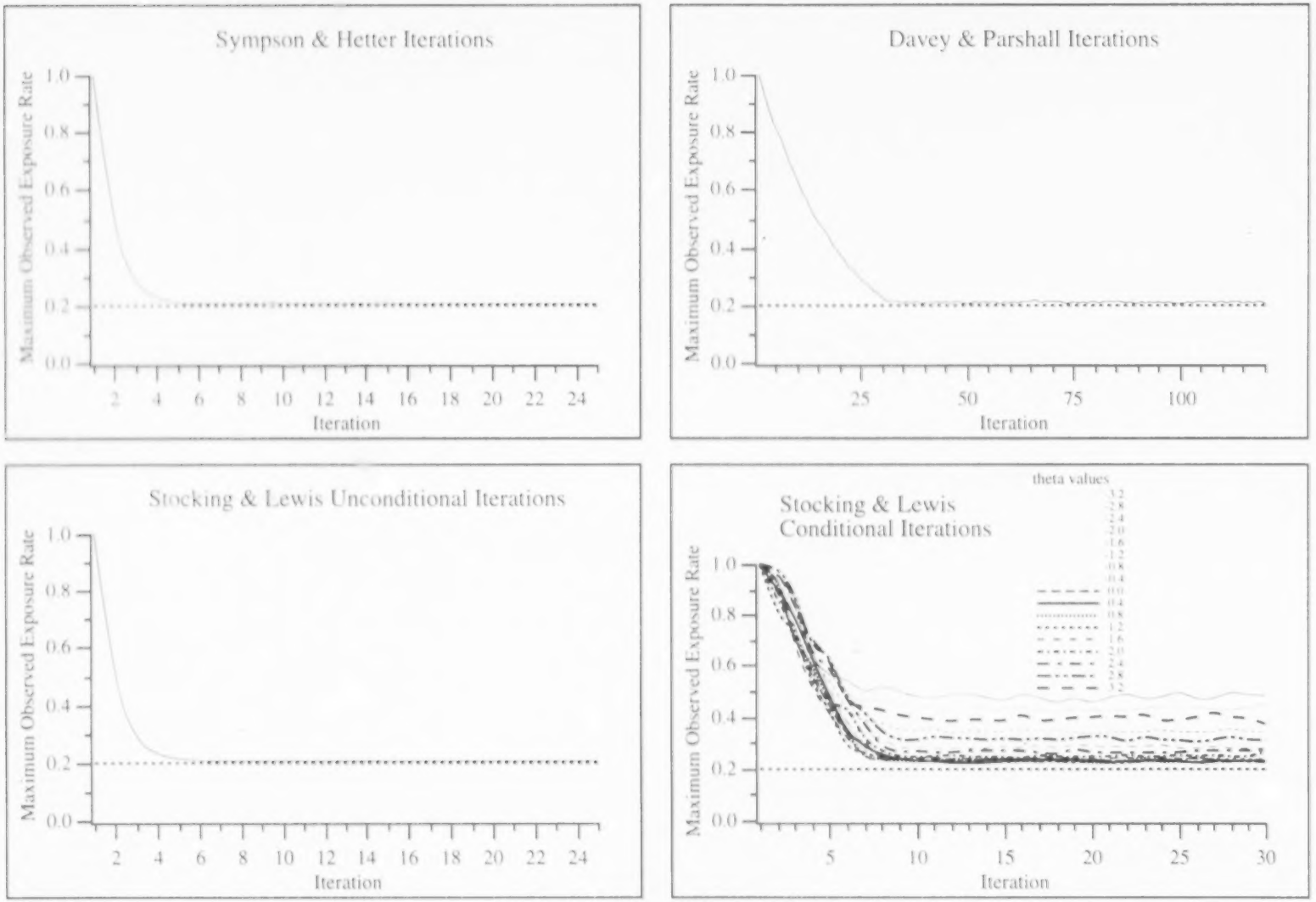

Figure 4. Results of Iterations for the Large Pool with r= 20

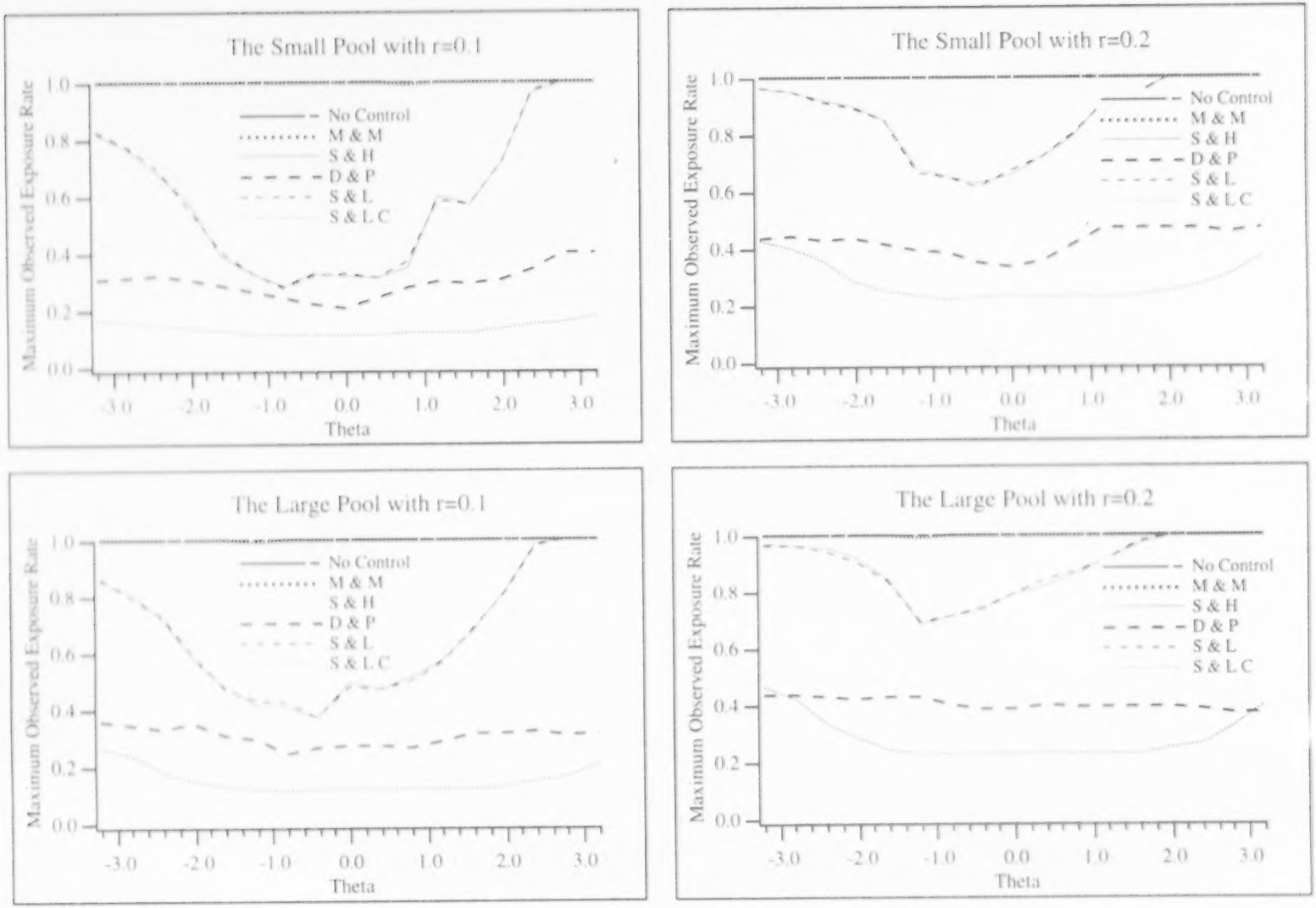

Figure 5. Conditional Maximum Observed Exposure Rates for the Various Conditions

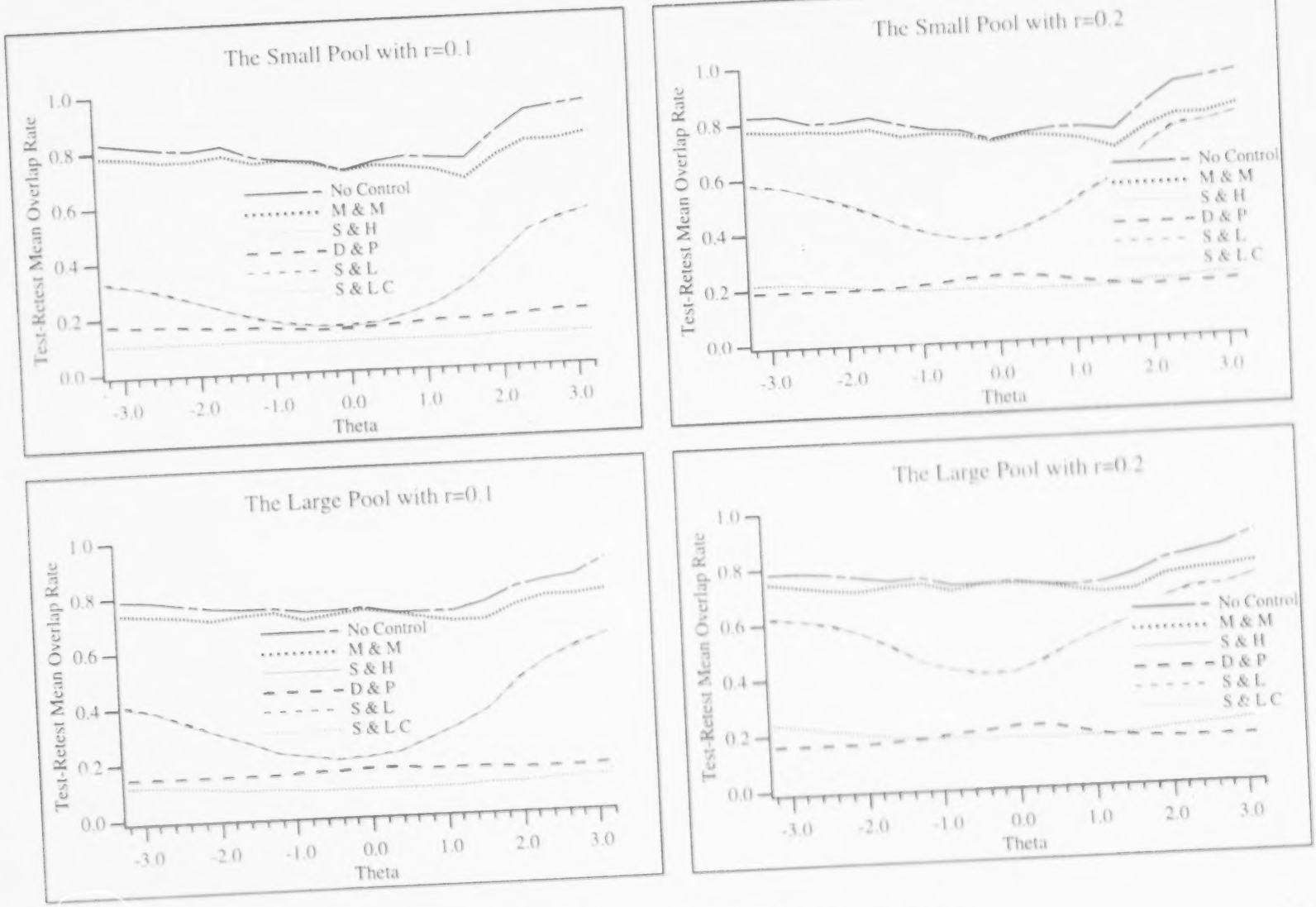

Figure 6. Test-Retest Mean Overlap Rates for the Various Conditions

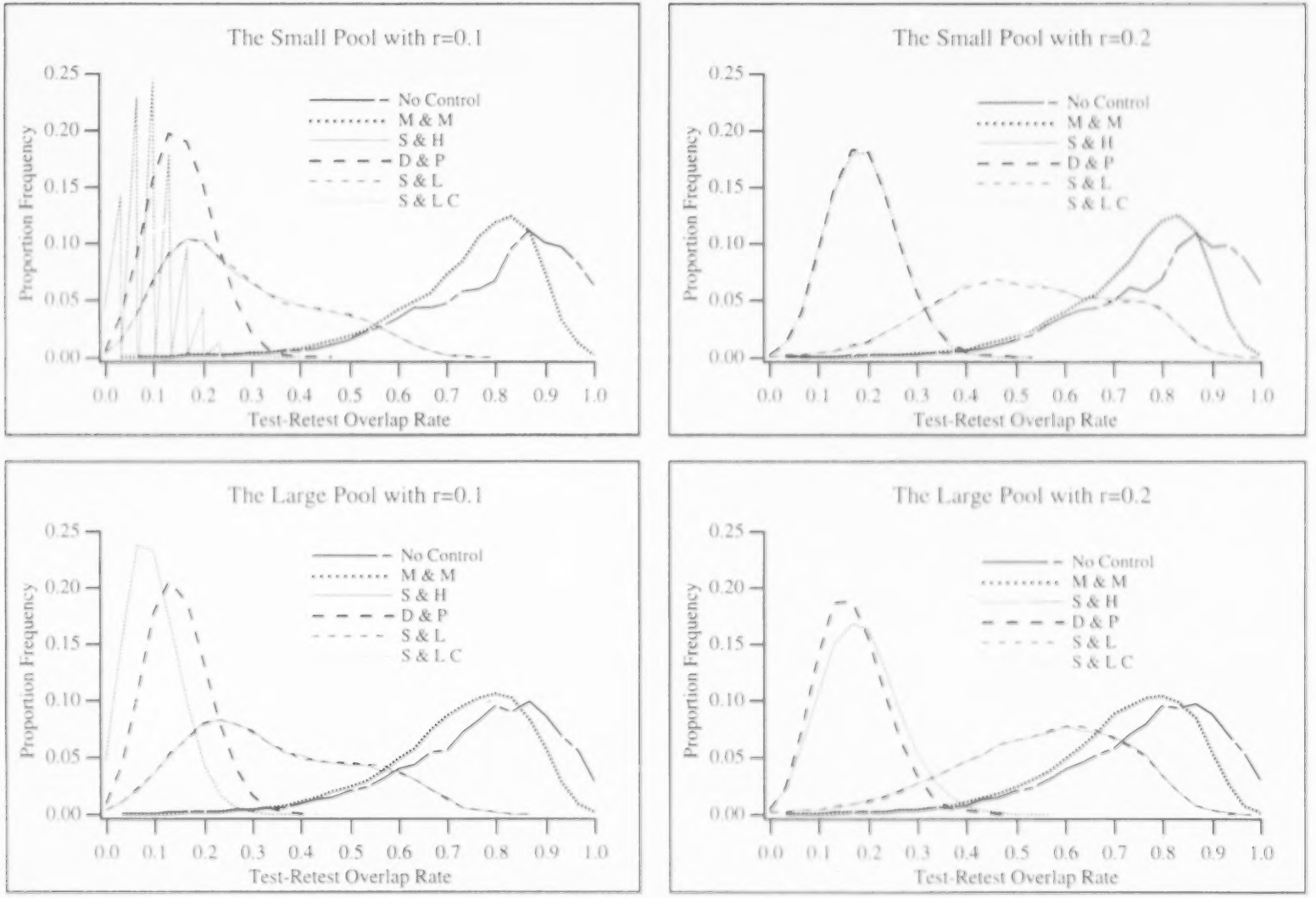

Figure 7. Full Distribution of the Test-Retest Overlap Rates for the Various Conditions

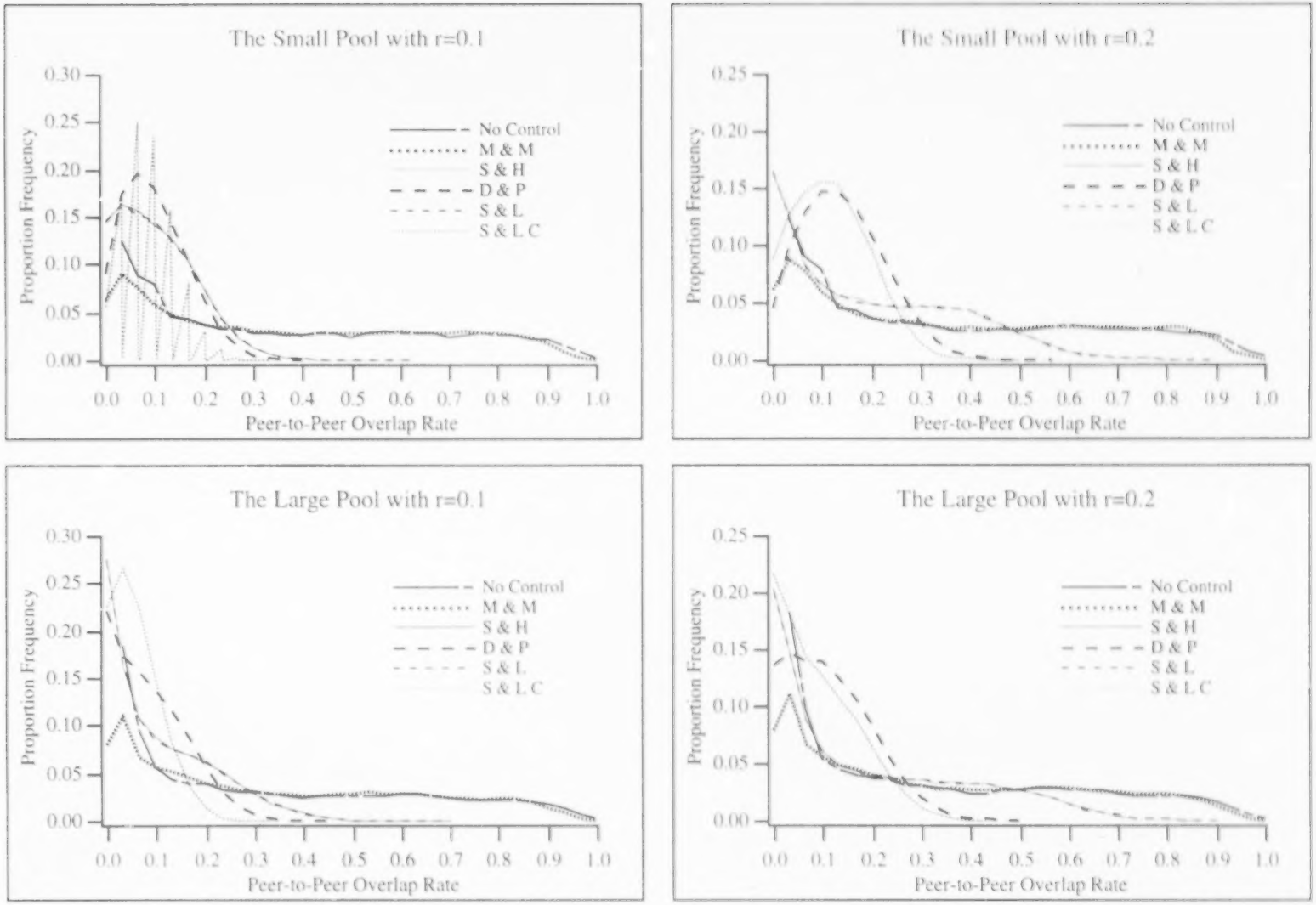

Figure 8. Full Distribution of the Peer-to-Peer Overlap Rates for the Various Conditions

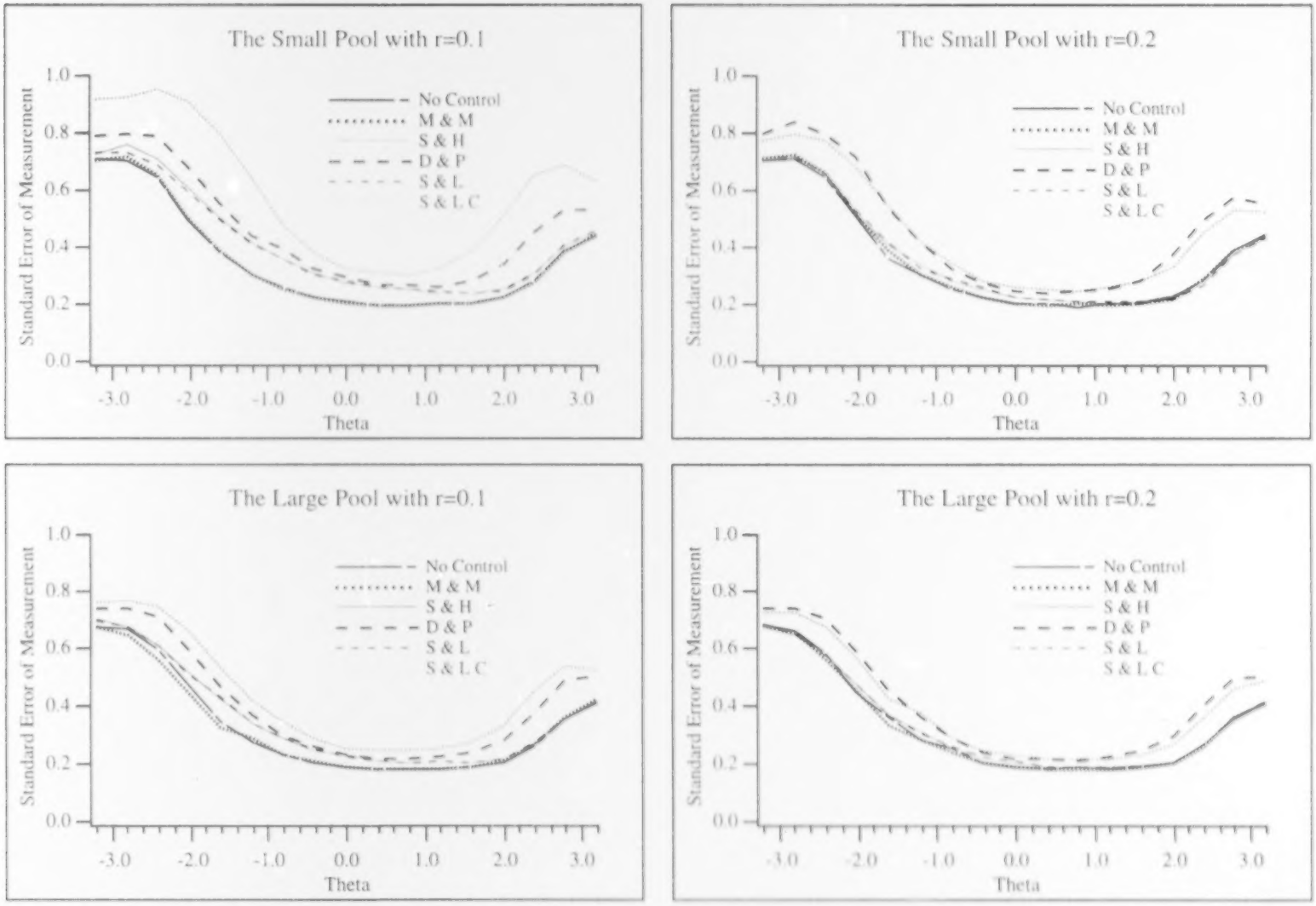

Figure 9. Standard Errors for the Various Conditions